D654–D657 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 7 November 2006*
doi:10.1093/nar/gkl808

PigGIS: Pig Genomic Informatics System

Jue Ruan[1,2], Yiran Guo[1,2], Heng Li[1], Yafeng Hu[1], Fei Song[1], Xin Huang[1], Karsten Kristiensen[3], Lars Bolund[1,4,]* and Jun Wang[1,3,4,]*

[1]Beijing Institute of Genomics of the Chinese Academy of Sciences, Beijing Genomics Institute, Beijing 101300, China, [2]Graduate University of the Chinese Academy of Sciences, Yuquan Road 19A, Beijing 100039, China, [3]Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense M, Denmark and [4]Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark

Received August 14, 2006; Revised September 21, 2006; Accepted September 29, 2006

ABSTRACT

Pig Genomic Information System (PigGIS) is a web-based depository of pig (*Sus scrofa*) genomic learning mainly engineered for biomedical research to locate pig genes from their human homologs and position single nucleotide polymorphisms (SNPs) in different pig populations. It utilizes a variety of sequence data, including whole genome shotgun (WGS) reads and expressed sequence tags (ESTs), and achieves a successful mapping solution to the low-coverage genome problem. With the data presently available, we have identified a total of 15 700 pig consensus sequences covering 18.5 Mb of the homologous human exons. We have also recovered 18 700 SNPs and 20 800 unique 60mer oligonucleotide probes for future pig genome analyses. PigGIS can be freely accessed via the web at http://www.piggis.org/ and http://pig.genomics.org.cn/.

INTRODUCTION

Besides being a source of food (1) the pig was known as an important model organism for evolutionary and biomedical research. It belongs to Artiodactyla, an order different from Rodentia and Primates. It stands at a unique position in recent comparative studies. Although in evolutionary history the pig, mouse and human species were separated at a similar time (2), the pig genome is much more similar to man than mouse (3) mainly due to the extensive genome rearrangements among the rodents (4). This fact makes pig an attractive model in biomedical studies—especially since the organ structures, physiology and metabolism also are very similar to those of man. Furthermore, the longevity and size of the pig allows longitudinal studies of chronic degenerative disease processes. Complex traits such as obesity and cardiovascular diseases are being extensively studied in pig models. At the same time, pig is also a very useful animal in the research fields of reproduction, tissue degeneration/biological maintenance, genetic modification/functional genomics, stem cell research and xenotransplantation (5,6).

There are many publicly available resources presenting information on porcine quantitative trait loci (QTL) mapping, and some other web services have been developed to provide genome level sequence data of the animal (4,7,8). Most of them only provide unigenes assembled from expressed sequence tags (ESTs). Genomic reads are available in the website of the Sino–Danish Pig Genome Project, but are left to be assembled or annotated due to the low sequencing coverage (0.66×). Given such rich pig resources, there is a need to combine them and produce a more complete and versatile annotation. This is the motivation of the Pig Genome Informatics System (PigGIS).

PigGIS aims to present the most complete pig genome annotation to date. It is based on the 0.66× genomic reads and ESTs generated by the Sino–Danish Pig Genome Project, but integrates all the pig sequences available in GenBank. Taking well-annotated human genes as templates, PigGIS establishes a cost-effective pipeline that accurately predicts the alignment of each piece of orthologous pig sequence to the human genes. PigGIS sets a good example of what can be achieved with low-coverage sequencing data, given the presence of closely related genomes.

DATA SOURCE

PigGIS is based on three categories of data. The first type of sequences are 3.84 million whole genome shotgun (WGS) reads generated by the Sino–Danish Pig Genome Project (4). The average trimmed length of these reads extends to 543 bp, yielding a total of 2.1 billion bp, which is equivalent to 0.66× coverage of the 3.15 Gb pig genome. The second type of data consists of 870 084 ESTs from 100 differentiated

*To whom correspondence should be addressed. Tel: +86 10 80481552; Fax: +86 10 80498676; Email: wangj@genomics.org.cn
*Correspondence may also be addressed to Lars Bolund. Tel: +45 89421675; Fax: +45 86123173; Email: bolund@humgen.au.dk

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors pig tissues/developmental stages, which are also part of the results from the Sino–Danish Project. Finally 589 996 genomic reads together with 570 773 mRNA sequences were extracted from GenBank (9). Since PigGIS was developed to infer pig genes based on human–pig homology, we employed human annotations [22 218 human CDS from Ensembl (10) v32] as reference to anchor the pig sequences.

DATA PROCESSING METHODS

A multi-step procedure was performed to anchor WGS reads onto the human genes. Initially, we BLASTXed (11) WGS reads to human transcripts, filtered the reads with low identities and threw out the ones that may match lineage-specific duplications in the human genome. As the next step, we masked repeating sequences by RepeatMasker (http://www.repeatmasker.org/) and further aligned the resultant sequences to human exons by cross_match in the PHRAP (http://www.phrap.org/) package. Then, we discarded unmatched regions of the remaining reads, assembled matched ones by PHRAP, and mapped the pig contigs and singlets to the corresponding human coding exons by FASTY (12–15). Thus, we obtained the final anchored WGS reads.

By virtue of our mapping strategy of assembling sequence fragments aligned to human coding exons instead of genes, the same pipeline described above can be directly applied to ESTs and mixtures of WGS reads and ESTs. Thus, ESTs obtained by screening were included in the final clusters.

Similar to our previous analysis on the chicken genomes (16), a search for single nucleotide polymorphisms (SNPs) was carefully carried out among the five pig strains through comparisons between high-quality base sequences residing in the WGS reads and ESTs. Two SNP sets were constructed with our pipeline. The first set was recovered directly from high-quality base pair difference, whereas the second by ruling out the SNPs in the first set which are too close to each other.

DATA CONTENTS

At present, the portion of known human coding exons covered by our final anchored pig WGS reads is 66%, corresponding to 14 618 human genes. It accounts for 27% of the size of all the human coding exons (i.e. a total span of 8 987 719 bp). The anchored pig ESTs are overlapped with 55% (12 299) of human genes, contributing to 26% (8 566 083 bp) of human coding sequences. If we take WGS reads and ESTs as a whole (defined as ‘both’ in some of the PigGIS web pages), our data cover ~73% (16 309) of the human genes, spanning 42% (13 681 830 bp) of the coding human genomic length. These numbers increase to 76% (16 958) of the genes, spanning 56% (18 548 727 bp) of the coding length, when additional 1.2 million miscellaneous GenBank records are taken into account (this explains the meaning of ‘all’ presented in several subviews of the system). Constructed to mine the genetic variations in different breeds of pigs, PigGIS also makes use of a systematic workflow to find SNPs. At present, a total of 18 712 SNPs have been located. Additionally, 20 786 unique 60mer oligonucleotide probes have been designed to facilitate future pig genome analyses. Table 1 incorporates these numerical data into a clear summary.

Table 1. Data summary

Data types	Counts[a]	Base pairs covered[b]
WGS reads		
Hampshire	707 281	—
Yorkshire	1 204 666	—
Landrace	650 609	—
Duroc	1 015 722	—
ErHuaLian	256 993	—
Total	3 835 271	—
ESTs	870 084	—
Homologous gene sequences (human versus pig)		
WGS covered	14 618 (66%)	8 987 719 (27%)
EST covered	12 299 (55%)	8 566 083 (26%)
Covered by both	16 309 (73%)	13 681 830 (42%)
+GenBank sequences	16 958 (76%)	18 548 727 (56%)

[a] Percentages of covered gene numbers are given in parentheses.
[b] Percentages of covered coding genomic length are given in parentheses.

ACCESS AND WEB QUERY INTERFACE

We designed two ways for users to access the PigGIS: browse and query. In order to fetch the information, one can browse by clicking on a specific region in the human genome, or one can submit a keyword inquiry about a human gene for detail. Take the *TNR* (human tenascin-R) gene for example. At present, its porcine counterpart is beyond immediate access in the NCBI nucleotide database. In order to obtain the ‘TNR’ in pig, traditionally we would have to use a complicated procedure consisting of searching for the sequences of the human ‘TNR’, fishing out the pig homologous fragments by BLAST, and then PHRAPping them to shape the pig ortholog. Now, PigGIS makes it an easy typing-and-clicking experience. One may find the ‘TNR’ by clicking 1q25.1 of the human chromosome 1 on the homepage, or more directly, by entering ‘TNR’ in the search box. After that, the TransView page (Figure 1) of TNR will provide details about the homologous pig clusters which cover 46% of the coding sequences of the human gene and contains one SNP. This ‘TNR’ example also manifests how PigGIS manipulates frameshifts. Searching for strings ‘TCAACC-’ and ‘GGAGGC-TCAGCT’ in the TransView page, the user may find a deletion and an insertion to the pig genome, respectively. Both of them can incur local frameshifts. We arbitrarily replace them with consensus sequences without frameshifts, shown at the bottom of the page. Although in this example the two frameshifts might be true, removing them in most other cases turns out to be a better choice than keeping them, at least to the best of our experience. In the TransView page, the user can follow the links in the sidebar to examine the raw sequence data of which the pig consensus consists, and also to view the expressed information given by ESTs, the contigs that contribute to the pig CDS, the oligonucleotide probes that uniquely represent the gene, and pig SNPs found in comparisons between pig sequences. Alternatively, the user can click on the short lines in the figure to see the various fragments and oligos,

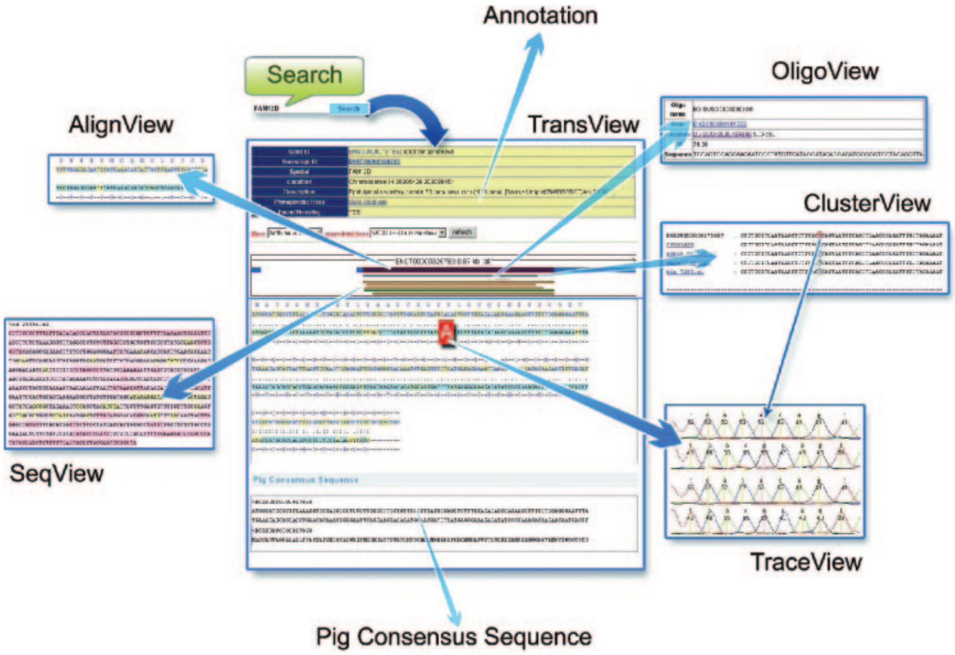

Figure 1. Screenshots of a virtual paradigm in PigGIS. Centered is the TransView which contains a group of informative components with the pig consensus sequence at the bottom. Top left is the AlignView, showing sequences of pig clusters aligned to exons in a human gene, and ranking them by their identity to the human sequences. Top right are the OligoView and ClusterView, presenting unique 60mer oligonucleotide probe information and visualizing the multiple pig sequences assembling the clusters, respectively. Bottom left is the SeqView providing the raw sequence contents of the reads/ESTs. Bottom right is the TraceView showing the raw evidence supporting the annotation of SNPs in the pig sequence.

and click highlighted nucleotides, if they exist, to carefully check SNPs. PigGIS provides detailed assembly of each contig, and displays qualities and chromatogram charts whenever possible, facilitating the validation of each base pair of the pig sequences.

SYSTEM IMPLEMENTATION

Evolving from our former database systems (16–18), PigGIS is composed of four hardware components: a World Wide Web server running Solaris operating system which is fully supported by Java Runtime Environment, a separated database server managing MySQL queries, a sequence analysis/homology search server and an FTP server for raw sequence downloads. Most of the web services are based on an Apache+TomCat architecture and was developed by JSP/Servlet/JavaBean technology. Also, Java Applets are employed to enable principle TraceView functions in the front end. The back end sequence analysis codes were written in PERL. Such hybrid style of programming has proven itself simple yet efficient.

DISCUSSION

Advantages of our strategy

Our pipeline anchors sequences before assembling, but does not assemble WGS reads or ESTs from scratch. This strategy requires much less computing resources, and also makes it possible to annotate the regions covered by WGS singlets which are usually discarded as sequencing errors. Although singlets are error-prone, to discard all of them is too conservative. By ruling out spurious hits in various ways, problems caused by singlets can be reduced to a minimum. Another distinct advantage of our method is that the pipeline can take both genomic and spliced sequences as input. WGS reads, BACs, ESTs and other public sequences are processed in almost the same way. They can be freely mixed together, which will lead to far better coverage of the coding sequences.

The approach we have chosen when constructing the pig consensus sequences might be relevant to the NIH-funded Mammalian Genome Project (http://www.broad.mit.edu/mammals/) underway, which is an effort aiming at producing low-coverage (2×) assemblies for 16 more mammals, in addition to the five completely sequenced ones, to achieve alignments for more extensive cross-species comparisons than before. Our approach, to some extent, implemented such a task at an ultralow-coverage level.

Known limitations

Our data covered less human sequences than previously presumed. We mainly ascribe this to four reasons: (i) desertion of shorter matches existing in human coding exons and WGS reads, (ii) absence of homologous sequences of pig genes in man, (iii) exclusion of hits anchored to recent duplications in the human genome, and finally (iv) failure of the fast evolving regions to pass our stringent criteria.

Although it would be possible to revise related criteria and get higher coverage, accuracy would probably be affected at the same time. Consequently, we still decided to stay with more conservative annotations that contained fewer errors.

Moreover, the combination of a variety of data types, including ESTs and miscellaneous GenBank sequences, offers some compensation that will keep the low-coverage problem from being too serious.

Another difficulty lurks in the lineage-specific duplications in the pig genome. This probably underlies some of the spurious SNPs. Consequently, we organized two sets of SNPs, with the smaller one containing fewer errors but missing some of the true SNPs in the bigger one.

Future developments

Continued efforts will be invested in collecting more pig genomic data, refining our annotation methods, improving the handling of the known problems listed above, and retrieving and integrating useful knowledge behind the pig genome sequences. We will also apply this kind of sequencing and annotating scheme to low-coverage genome information from other mammals that are evolutionarily close to man.

ACKNOWLEDGEMENTS

This project was supported by Chinese Academy of Sciences (GJHZ0518), Ministry of Science and Technology under program CNGI-04-15-7A, Ministry of Education (XXBKYHT2006001), National Natural Science Foundation of China (90608010; 90208019; 90403130; 30221004; 90612019), and China National Grid. The work was further supported by the Danish Platform for Integrative Biology, the Ole Rømer grant from the Danish Natural Science Research Council and the Danish Medical Research Council. Funding to pay the Open Access publication charges for this article was provided by the Danish Medical Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Fang,M., Hu,X., Jiang,T., Braunschweig,M., Hu,L., Du,Z., Feng,J., Zhang,Q., Wu,C. and Li,N. (2005) The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Animal Genet*., **36**, 7–13.
2. Jorgensen,F.G., Hobolth,A., Hornshoj,H., Bendixen,C., Fredholm,M. and Schierup,M.H. (2005) Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. *BMC Biol*., **3**, 2.
3. Yu,Z., Li,Y., Meng,Q., Yuan,J., Zhao,Z., Li,W., Hu,X., Yan,B., Fan,B., Yu,S. *et al*. (2005) Comparative analysis of the pig BAC sequence involved in the regulation of myostatin gene. *Sci. China*, **48**, 168–180.
4. Wernersson,R., Schierup,M.H., Jorgensen,F.G., Gorodkin,J., Panitz,F., Staerfeldt,H.H., Christensen,O.F., Mailund,T., Hornshoj,H., Klein,A. *et al*. (2005) Pigs in sequence space: a 0.66× coverage pig genome survey based on shotgun sequencing. *BMC Genomics*, **6**, 70.
5. Rothschild,M.F. (2003) From a sow's ear to a silk purse: real progress in porcine genomics. *Cytogenet. Genome Res*., **102**, 95–99.
6. Vodicka,P., Smetana,K., Jr, Dvorankova,B., Emerick,T., Xu,Y.Z., Ourednik,J., Ourednik,V. and Motlik,J. (2005) The miniature pig as an animal model in biomedical research. *Ann NY Acad. Sci*., **1049**, 161–171.
7. Fadiel,A., Anidi,I. and Eichenbaum,K.D. (2005) Farm animal genomics and informatics: an update. *Nucleic Acids Res*., **33**, 6308–6318.
8. Uenishi,H., Eguchi,T., Suzuki,K., Sawazaki,T., Toki,D., Shinkai,H., Okumura,N., Hamasima,N. and Awata,T. (2004) PEDE (Pig EST Data Explorer): construction of a database for ESTs derived from porcine full-length cDNA libraries. *Nucleic Acids Res*., **32**, D484–D488.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res*., **34**, D16–D20.
10. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
12. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad Sci. USA*, **85**, 2444–2448.
13. Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol*., **266**, 227–258.
14. Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
15. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol*., **132**, 185–219.
16. Wang,J., He,X., Ruan,J., Dai,M., Chen,J., Zhang,Y., Hu,Y., Ye,C., Li,S., Cong,L. *et al*. (2005) ChickVD: a sequence variation database for the chicken genome. *Nucleic Acids Res*., **33**, D438–D441.
17. Wang,J., Xia,Q., He,X., Dai,M., Ruan,J., Chen,J., Yu,G., Yuan,H., Hu,Y., Li,R. *et al*. (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res*., **33**, D399–D402.
18. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al*. (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res*., **32**, D377–D382.